Computer Assisted Detection and Analysis of Tall Cell Variant Papillary Thyroid Carcinoma in Histological Images

Edward Kim[1], Zubair Baloch[2], Caroline Kim[3]

[1] Department of Computing Sciences, Villanova University, Villanova, PA;
[2] Department of Pathology and Laboratory Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA;
[3] Division of Endocrinology, Diabetes, and Metabolism, Department of Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA;

ABSTRACT

The number of new cases of thyroid cancer are dramatically increasing as incidences of this cancer have more than doubled since the early 1970s. Tall cell variant (TCV-PTC) papillary thyroid carcinoma is one type of thyroid cancer that is more aggressive and usually associated with higher local recurrence and distant metastasis. This variant can be identified through visual characteristics of cells in histological images. Thus, we created a fully automatic algorithm that is able to segment cells using a multi-stage approach. Our method learns the statistical characteristics of nuclei and cells during the segmentation process and utilizes this information for a more accurate result. Furthermore, we are able to analyze the detected regions and extract characteristic cell data that can be used to assist in clinical diagnosis.

Keywords: cell segmentation, cell analysis, papillary thyroid carcinoma, tall cell variant

1. INTRODUCTION

Thyroid cancer is the fastest increasing cancer in both men and women and is the most common endocrine cancer.[1] According to the National Cancer Institute's Surveillance, Epidemiology, and End Results Program (SEER),[2] there will be 62,980 new cases of thyroid cancer and 1,890 people will die of this disease in 2014. Additionally, thyroid cancer cases have been rising on average 5.5% each year over the last 10 years. Papillary and follicular thyroid carcinomas are well-differentiated and account for 80-90% of all thyroid cancers. However, within papillary cancer, there exist several variants that have been described based upon their morphological features and clinical behavior. These variants ultimately can have a significant impact on recurrence and mortality. One variant, tall cell variant (TCV-PTC), is an aggressive papillary thyroid carcinoma that is more likely to be present with extra thyroidal extension. TCV-PTC has an incidence rate between 3.8% - 10.4% and has shown to have a 4.5x recurrence rate and 14.28x increase in disease mortality when compared to the classical, usual variant of papillary thyroid carcinoma (UV-PTC).[3] Additionally, the tall cell variant of papillary thyroid carcinoma has a poorer 5-year disease survival (81.9% vs 97.8%).[4] Thus, it is important that the diagnosis of TCV-PTC be accurate and precise.

2. BACKGROUND

Unfortunately, accurate diagnosis of TCV-PTC is a challenging problem for several reasons. One issue is with the characterization of TCV. Hawk and Hazard were the first to identify this subtype in 1976, stating that the tall cell variant is characterized by tumor cells that are at least twice as long as they are wide.[5] Recent literature from the World Health Organization has updated this characteristic to cells being three times their width.[6] The change in guidelines can cause ambiguity and further raises the question as to the significance of the length-width ratio and the effect of this ratio on patient outcomes. Specifically, what is the correlation (or is there even a correlation) between the length-width cell ratio and recurrence, presence of extra thyroidal extension, and mortality? It is unknown how this factor *independently* relates to the prognosis.

The second challenge is with the classification of a TCV tumor. The literature has diagnosed the TCV in tumors where the TCV comprised anywhere from 10% to 70% of a particular tumor.[7] More recent literature

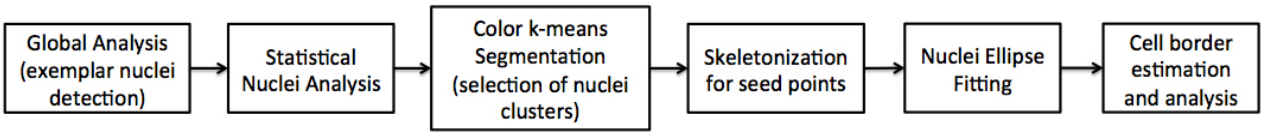


Figure 1. Flowchart describing the steps of our method. A global analysis of the image is performed to obtain exemplar nuclei. This is followed by a color, size, and shape analysis of the nuclei. k-means segmentation is performed and clusters belonging to nuclei and background are identified. The segmentation is skeletonized and the branch points are used as seeds for a least squares ellipse fitting method. Finally, the cell borders can be estimated for a cell analysis.

has suggested that TCV-PTC should comprise of at least 50% morphology.[7–9] Similar to the length-width ratio, there is a threshold number that can be applied to the classification of a tumor; however, there is no direct correlation data between the percent composition of a tumor and outcome.

These complications are now being brought to the forefront when performing a pathological review.[10] "How much of a tumor needs to show the features of TCV to be diagnosed as such?" "What is not TCV?" And most importantly to a patient, "What are the consequences of a diagnosis of TCV?"

As a first step towards addressing some of these questions, we propose a method that detects, analyzes, and describes the presence of the tall cells in histological images. Our method is fully automatic and is able to generate a statistical model of the nuclei and cells within images during the segmentation process. We are then able to count the number of regular cells, tall cells, and furthermore analyze the cells with regard to their length-width ratio, percent of tall cells in the image, and cytological features, and nuclear crowding. As a guide for the cytological criteria, we reference the work by Solomon et al.[11] This study assessed the presence of seven cytologic features in TCV and UV-PTC. These features are, papillary groups, elongated/tall cells, oncocytic cytoplasm, distinct cell borders, prominent central nucleoli, nuclear grooves, and intranuclear inclusions. The conclusion of this work states that cytologic features can distinguish TCV from UV-PTC and the most significant features were elongated/tall cells, oncocytic cytoplasm, distinct cell borders, and "soap-bubble"-like intranuclear inclusions.

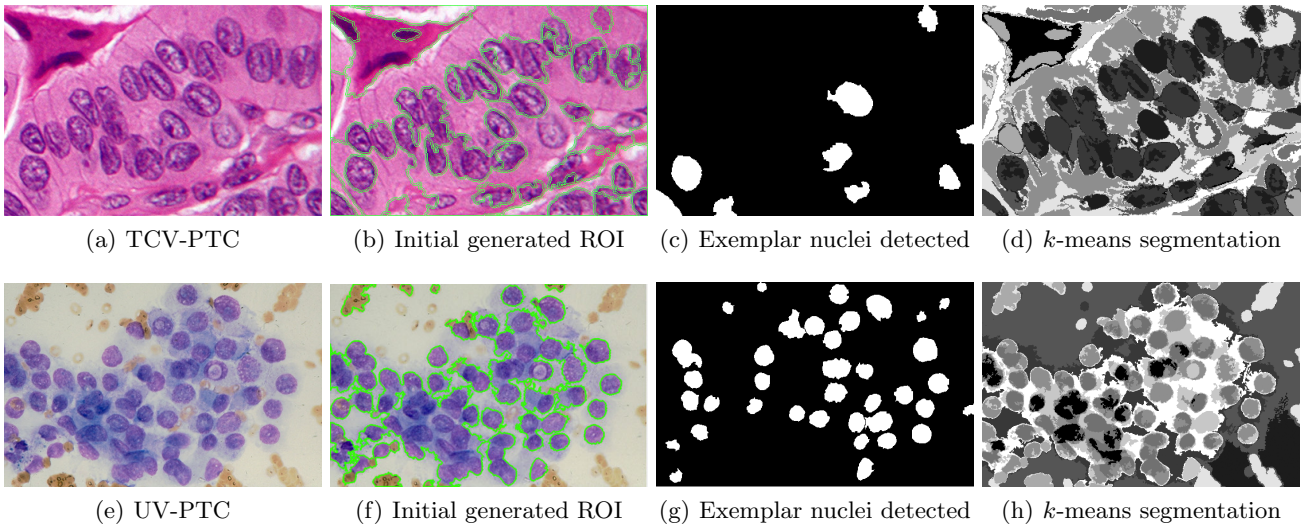
Our method is primarily aimed at identifying the elongated/tall cell feature of papillary carcinoma; however, it is well suited to be extended for additional classification. Qualitative and quantitative results demonstrate the effectiveness of our method in detecting and describing tall cells. Our long term goal is to have this program be an adjunctive tool in TCV-PTC diagnosis, as well as serve as a data collection mechanism for a larger regression study on the effects of cell morphology on patient outcomes.

3. METHODOLOGY

Our method is a multi-stage image segmentation algorithm. A flowchart of the steps of our algorithm can be seen in Figure 1. The fundamental idea behind our framework is to begin with a high level analysis of the input image, gather global statistics of the image and nuclei, and then perform precise nuclei segmentation using both global and local image data. The nuclei will be used in a cell estimation technique. The stages of our framework are described in detail below.

3.1 High Level, Global Image Analysis

In the first step, we perform a high level, global analysis of the image. A basic thresholding is performed, along with morphological opening and closing of the binary image in order to find possible exemplar nuclei that can be used for training a statistical model of what characterizes a nucleus. A watershed algorithm is applied to the modified thresholded image and potential regions of interest are created, see Figure 2(b)(f). Each area is analyzed for convexity, and only the areas that meet a high convexity threshold (>80%) and eccentricity threshold (<80%) are selected as exemplar nuclei, see Figure 2(c)(g). The image is converted from the RGB color space to $L^*A^*B^*$ space and statistics are gathered from the exemplar cells including, average cell pixel area, standard deviation of area, average $L^*A^*B^*$ color, and associated color standard deviation in each channel. In the event that no exemplar cells can be identified automatically, our method allows for a manual selection of cells for initial training.



(a) TCV-PTC (b) Initial generated ROI (c) Exemplar nuclei detected (d) k-means segmentation

(e) UV-PTC (f) Initial generated ROI (g) Exemplar nuclei detected (h) k-means segmentation

Figure 2. Visualization of the high level, global pass of the algorithm. The steps of visualized on a TCV-PTC image can be seen in (a)-(d). The steps visualized on a UV-PTC image can be seen in (e)-(h). The k-means segmentation is performed with k=10 classes.

With the statistical model of what a cell nucleus looks like in the given test image, we can perform a 3 dimensional L*A*B* color k-means segmentation, $k = 10$, on the image, see Figure 2(d)(h). We can then use the exemplar nuclei color model to identify which k segments correlate with the cells, and distinguish between the foreground (cells) and background. Specifically, this correlation is performed by computing the 3D Euclidean distance between cluster centers and the mean L*A*B* color of the exemplar nuclei. The resulting k-means segmentation is used to identify the preliminary segmentation for the next stage of our algorithm, see Figure 3(a)(e), that provide the binary image regions used in the individual nuclei segmentation.

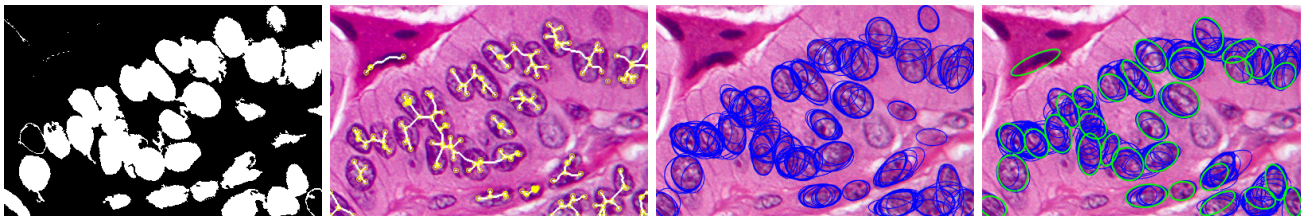
3.2 Individual Nuclei Segmentation

In this next step, we analyze the given k-means segmentation and identify and analyze the nuclei that match the color and size of exemplar nuclei. Unfortunately, with most variants of PTC, cells can exhibit nuclear crowding, where the nuclei borders touch, or even overlap each other. An example of overlapping nuclei that create a merged region of interest can be seen in Figure 3(e) where there is a large binary region of white that encompasses more than one nuclei. These regions can be mathematically identified by searching for area regions that are beyond two standard deviations of a typical nuclei and are within a 3D Euclidean color distance threshold, ($t = 40$) of the nucleus model. In order to properly segment overlapping cells, we modify an ellipse fitting segmentation algorithm, SNEF,[12] to work with our application.

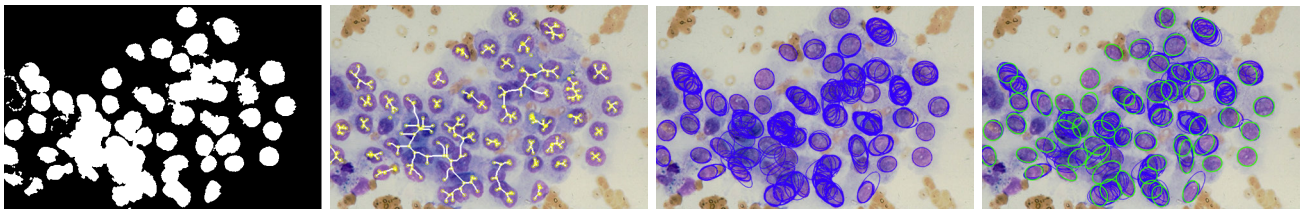
3.2.1 Modified Ellipse Fitting for Nuceli Segmentation

A modified SNEF method[12] is the basis for this step of the segmentation framework. We briefly summarize the method here, and emphasize the modifications we proposed with our algorithm. First, we morphologically open and then skeletonize the region of interest to get centroid points as seed points, s_x, s_y, for an ellipse fitting method, see Figure 3(b)(f). This skeletonization procedure removes pixels on the boundaries of objects but does not allow objects to break apart. The pixels remaining make up the image skeleton. The set of seeds for our framework consists of the set of branch points of the skeleton and the set of edge points. In contrast to Hukkanen et al.[12] which uses ultimate erosion to get seed points, empirically, the skeletonization method provides fewer candidates and more accurate seed points.

For each seed point, we extend a ray outwards in all angles, $\alpha \in \{1°, \ldots, 360°\}$. A set of points on the ray can be created by using the formulation, $x_i = s_x + r\cos\alpha$ and $y_i = s_y + r\sin\alpha$. The length of the ray, r, is extended ($r = r + 1$) until the ray intersects the boundaries of the k-means segmentation. The intersection points will create a set of connected component segments, $C_1 \ldots C_n$, that surround the seed point. The connected



(a) k-means cluster selection (b) Skeleton with seed points (c) Candidate ellipses (d) Final ellipses



(e) k-means cluster selection (f) Skeleton with seed points (g) Candidate ellipses (h) Final ellipses

Figure 3. Individual nuclei segmentation steps of the algorithm. The steps of visualized on a TCV-PTC image can be seen in (a)-(d). The steps visualized on a UV-PTC image can be seen in (e)-(h). The yellow circles in (b) and (f) indicate both the branch and end points of the skeletonization method. The green ellipses in (d) and (h) are the final ellipses selected by the maximization of the goodness of fit function.

components are sorted by ascending Euclidean distance from the centroid point and the seed point. One by one these components are added to a list, $\mathcal{D}$, and an ellipse is fitted to the resulting point set. The ellipse parameters are estimated by a direct least squares fitting method.[13] We store every candidate ellipse, $e_1 \dots e_n$ into a set, $\mathcal{E}$, and compute find the best "fitting" ellipse to the image. The pixels representing a particular ellipse are defined as $\mathcal{P}(e, s_x, s_y)$.

3.2.2 Goodness of Fit Maximization

In order to evaluate the best fit of an ellipse to the image, we define a goodness of fit function. The goodness of fit takes into account the image gradient, the connected components obtained through raycasting, and the statistical characteristics of the cell nuclei. The fit of an ellipse can be calculated by the function, g, and the best fit is the ellipse $e_i \in \mathcal{E}$ that maximizes the goodness of fit. This is formally defined as,

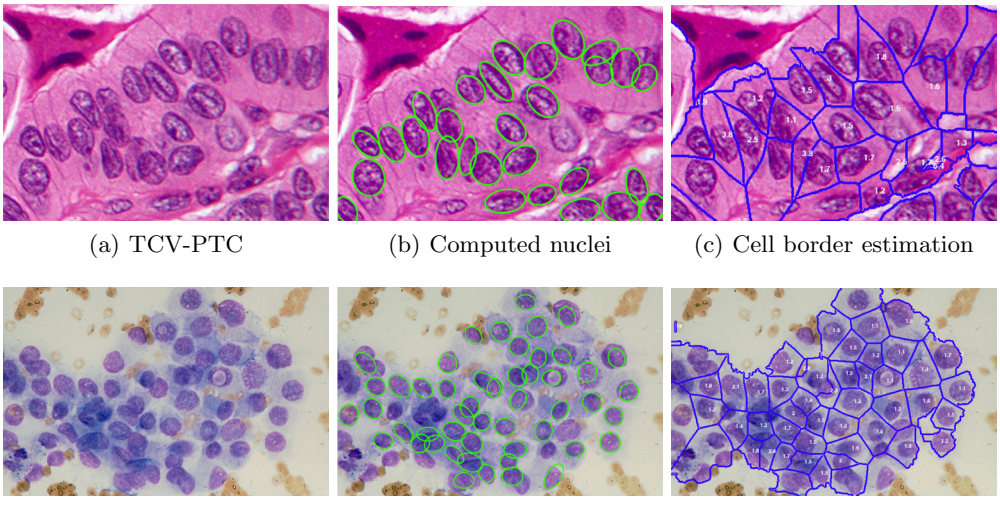
$$\arg\max_i g(i) = \lambda_1 \frac{|P(e_i, s_x, s_y) \cap \mathcal{H}|}{|P(e_i, s_x, s_y)|} + \lambda_2 \frac{|D_i \cap P'(e_i, s_x, s_y)|}{|D_i|} + \lambda_3 (1 - \frac{|A(e_i) - \bar{A}|_1}{2\bar{A}}) + \lambda_4 (1 - \frac{|T(e_i) - \bar{T}|_1}{2\bar{T}}) \quad (1)$$

The first term, computes the overlap of the candidate ellipse and $\mathcal{H}$, the dilated canny edges of the image. The second term computes the overlap of the dilated candidate ellipse ($\mathcal{P}'(e_i, s_x, s_y)$) and the set of points added to the connected component list up until, and including, segment D_i. The third term computes the normalized L1 norm of the difference between the candidate ellipse area and the mean area of the exemplar nuclei in the image, defined as $A(e_i)$ and $\bar{A}$ respectively. The last term computes the normalized L1 norm of the difference between the eccentricity of the candidate ellipse and the mean eccentricity of the exemplar nuclei, defined as $T(e_i)$ and $\bar{T}$ respectively. These last two terms are rewarding candidate ellipses that have similar shapes and sizes as the exemplar nuclei.

The candidate ellipses with areas and major-minor axis ratios greater than 4x their respective means are automatically discarded and not considered by this function. Furthermore, the λ's control the relative weights of each of the function components and are determined empirically in our framework. The following λ's are used, $\lambda_1 = 1$, $\lambda_2 = 3$, $\lambda_3 = 0.3$, and $\lambda_4 = 0.3$.

3.3 Cell Segmentation

The final stage of our framework involves the computation of the cell borders. In many cases of PTC, distinct cell borders do not exist and image gradient based methods will not accurately detect the boundary of cells.



(a) TCV-PTC (b) Computed nuclei (c) Cell border estimation

(d) UV-PTC (e) Computed nuceli (f) Cell border estimation

Figure 4. Visualization of the cell estimation method through a modified voronoi tessellation. The final cell borders in (c)(f) are highlighted in blue, and the computed length-width ratios are overlaid in the image.

Visually, one can see in Figure 2(e), the nuclei are prominently displayed; however, there are no cell borders present in the image. Thus we developed a cell border estimation technique based upon the nuclei position and background classification through the k-means segmentation.

We can estimate the cell through a modified voronoi diagram. Voronoi diagrams are a partitioning of spaces such that the area within a certain space is closer to a specified seed point than to any other. In a traditional voronoi diagram, the seed points form a set of coordinates. We modify the seeds so that our nuclei ellipses are used as seed points, from which a distance transform can be computed across the image. Background areas are masked from the distance transform and a watershed method is performed on the resulting combined distance map and mask. The result of this method serves as our estimated cell segmentation, see Figure 4(c)(f). This method works well with or without the existence of distinct cell borders. In our future work, we plan on refining the estimation in the event that these strong cell border cues do exist.

4. EXPERIMENTS AND RESULTS

For our experiments, we collected a total of 24 images of papillary thyroid carcinoma, 12 classified as classical UV-PTC, and 12 classified as TCV-PTC. Several images have been obtained from the Department of Pathology and Laboratory Medicine at the Perelman School of Medicine at the University of Pennsylvania, and others have been obtained from various research and educational thyroid cancer publications. On the collected images, we perform our cell segmentation algorithm and compare our method to several other cell segmentation algorithms, including an improved watershed algorithm[14] and a GMM based Hidden Markov Random Field method.[15] We present both qualitative and quantitative results that demonstrate the effectiveness of our method.

4.1 Qualitative Results

We present several qualitative results on TCV-PTC, see Fig. 5. Of particular significance is the ability of our system to handle nuclear crowding and overlapping cell and nuclei structures. Other methods typically merge overlapping regions together. Another advantage of our method is the ability to filter out circular regions that do not match the color of the exemplar cells. This has the benefit of eliminating false positives that show up in other methods, see Figure 5(b).

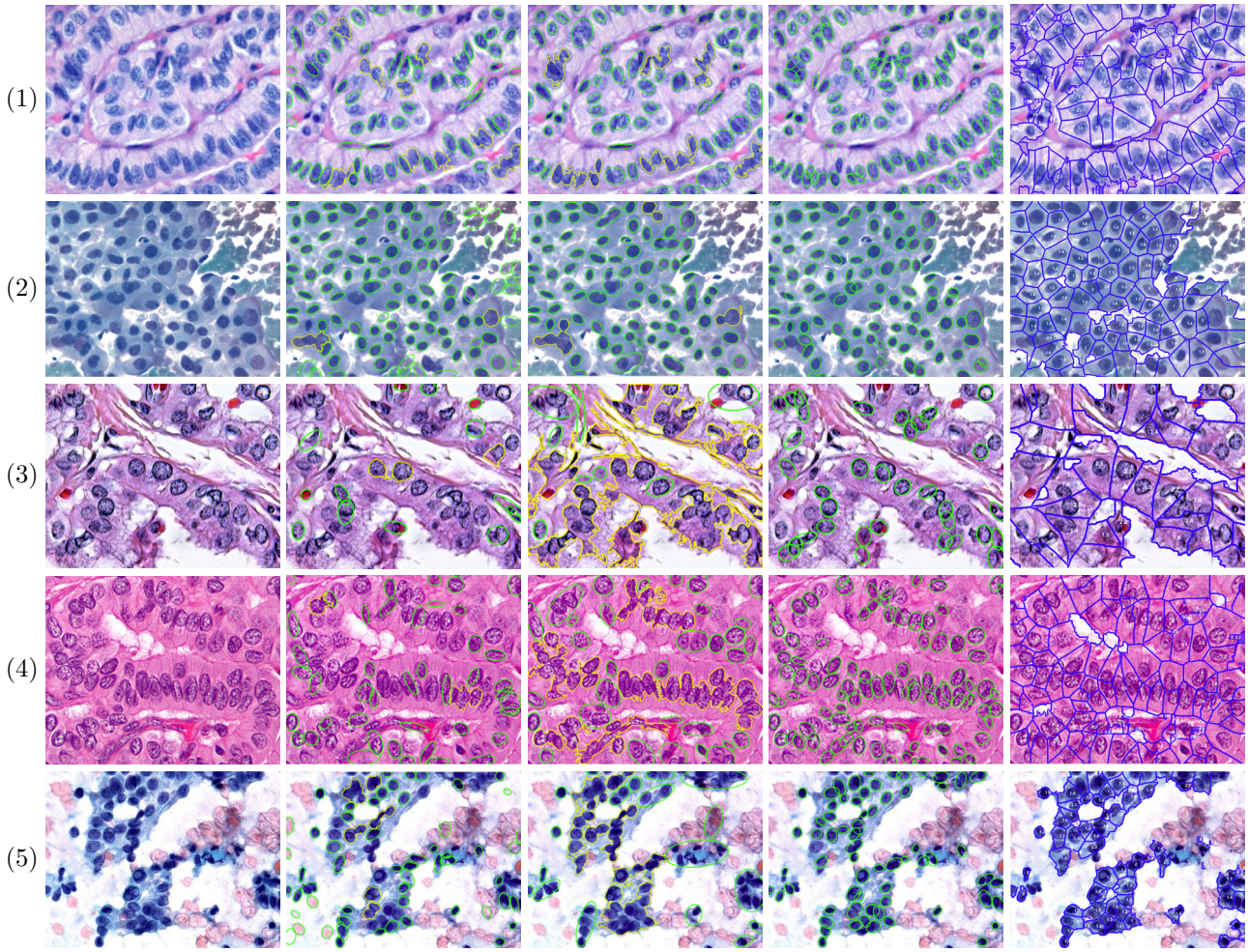


Figure 5. Segmentation and analysis results on TCV-PTC images. The green circles indicate detected and analyzed nuclei by the algorithm. The yellow lines in (b) & (c) indicate regions segmented by the method, but could not be analyzed due to irregular shape and size. Our ellipse fitting method is shown in (d). The final cell segmentation is shown in (e) with computed borders highlighted in blue.

4.2 Quantitative Results

We perform a quantitative analysis on the percent of cells nuclei segmented from classical PTC, and TCV-PTC, see Table 1. The ground truth is provided by manual annotation. Again, we are able to see through the quantitative results that our nuclei localization is fairly robust and accurate as compared to other methods. We attribute most of the success in our ability to segment overlapping structures.

For the nuclei that are correctly localized, we do a further quantitative analysis of our cell segmentation method. Ten percent of the cells are randomly selected and analyzed against manual ground truth to measure the accuracy of the length-width cell ratio. For TCV-PTC, we are able to compute the difference of the length-width ratio to the ground truth (where closer to zero difference is ideal) with a ratio accuracy of 0.506 ± 0.657. For UV-PTC we can compute the length-width ratio with an accuracy of 0.310 ± 0.262. This result is consistent with the nature of TCV-PTC versus UV-PTC as the tall cell variant has a much more variable length to width ratio.

Method	Type	% acc. nuclei	Type	% acc. nuclei
Imprv. Watershed[14]	UV-PTC	58.82%	TCV-PTC	65.98%
GMM-HMRF[15]	UV-PTC	49.74%	TCV-PTC	52.59%
Our Method	UV-PTC	**77.94%**	TCV-PTC	**83.78%**

Table 1. Accuracy of several methods when counting nuclei, UV-PTC and TCV-PTC. The percentage is calculated by comparing the output of the method and manually annotated ground truth.

5. CONCLUSION AND FUTURE WORK

Our segmentation and analysis framework for TCV-PTC is the first stage in our computerized understanding of TCV disease pathology. Our method automatically segments nuclei and cells in histological images of papillary carcinoma. We first find the exemplar nuclei and use these nuclei as a guide for finding the remaining regions of interest in an image. With the localized nuclei, our method finds the approximate cell borders by a modified voronoi diagram and analyzes the cells for their length-width ratio.

A recent study assessing 43,738 patients has found a 158% increased incidence of TCV diagnosis, as compared to a 60.8% increase in classical PTC.[10] Thus, it is increasingly important that automatic, computerized methods are available to support the clinical decision making process. Additionally, given an accurate and granular cell analysis, our future work will look at additional statistical analysis and regression on patient prognosis to find a mathematical model that might be able to correlate cell characteristics to patient outcome.

REFERENCES

[1] Hughes, D. T., Haymart, M. R., Miller, B. S., Gauger, P. G., and Doherty, G. M., "The most commonly occurring papillary thyroid cancer in the united states is now a microcarcinoma in a patient older than 45 years," *Thyroid* **21**(3), 231–236 (2011).

[2] Howlader, N., Noone, A., Krapcho, M., et al., "Seer cancer statistics review, 1975-2011.[based on the november 2013 seer data submission, posted to the seer web site, april 2014.]," *Bethesda, MD: National Cancer Institute* (2013). `http://seer.cancer.gov/statfacts/html/thyro.html`.

[3] Jalisi, S., Ainsworth, T., and LaValley, M., "Prognostic outcomes of tall cell variant papillary thyroid cancer: a meta-analysis," *Journal of thyroid research* **2010**(325602) (2010).

[4] Morris, L. G., Shaha, A. R., Tuttle, R. M., Sikora, A. G., and Ganly, I., "Tall-cell variant of papillary thyroid carcinoma: a matched-pair analysis of survival," *Thyroid* **20**(2), 153–158 (2010).

[5] Hawk, W. A. and Hazard, J. B., "The many appearances of papillary carcinoma of the thyroid," *Cleveland Clinic Quarterly* **43**(4), 207–216 (1976).

[6] DeLellis, R., Lloyd, R., Heitz, P., and Eng, C., "Pathology and genetics of tumours of endocrine organs. WHO classification of tumours," *The International Agency for Research on Cancer, Lyon, France* (2004).

[7] Johnson, T. L., Lloyd, R. V., Thompson, N. W., Beierwaltes, W. H., and Sisson, J. C., "Prognostic implications of the tall cell variant of papillary thyroid carcinoma," *The American journal of surgical pathology* **12**(1), 22–27 (1988).

[8] Akslen, L. A. and LiVolsi, V. A., "Prognostic significance of histologic grading compared with subclassification of papillary thyroid carcinoma," *Cancer* **88**(8), 1902–1908 (2000).

[9] Ghossein, R. and Livolsi, V. A., "Papillary thyroid carcinoma tall cell variant," *Thyroid* **18**(11), 1179–1181 (2008).

[10] Baloch, Z., LiVolsi, V. A., and Tondon, R., "Aggressive variants of follicular cell derived thyroid carcinoma; the so called real thyroid carcinomas," *Journal of clinical pathology* **66**(9), 733–743 (2013).

[11] Solomon, A., Gupta, P. K., LiVolsi, V. A., and Baloch, Z. W., "Distinguishing tall cell variant of papillary thyroid carcinoma from usual variant of papillary thyroid carcinoma in cytologic specimens," *Diagnostic cytopathology* **27**(3), 143–148 (2002).

[12] Hukkanen, J., Hategan, A., Sabo, E., and Tabus, I., "Segmentation of cell nuclei from histological images by ellipse fitting," *Proc. of the European Signal Processing Conference* , 1219–1223 (2010).

[13] Fitzgibbon, A., Pilu, M., and Fisher, R. B., "Direct least square fitting of ellipses," *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **21**(5), 476–480 (1999).

[14] Grau, V., Mewes, A., Alcaniz, M., Kikinis, R., and Warfield, S. K., "Improved watershed transform for medical image segmentation using prior information," *IEEE Transactions on Medical Imaging* **23**(4), 447–458 (2004).

[15] Wang, Q., "Gmm-based hidden markov random field for color image and 3d volume segmentation," *arXiv preprint arXiv:1212.4527* (2012).